Yashwanth Reddy Boddireddy

AI Engineer, Jersey City, NJ
+1 (848) 328 0485 | yashwanthreddy3047@gmail.com | Linkedin: yashwanth-reddy-boddireddy | GitHub: yashwanthreddy7178

Work Experience

AI Engineer | Enterprise AI Platforms Jan 2025 – Present
JPMorgan Chase & Co. | Remote, USA

- Designed and deployed scalable AI inference platforms using Docker, Kubernetes, and AWS EC2 GPU clusters, enabling 60K+ real-time inferences/ day, achieving 99.8% uptime and reducing model latency by 35% for enterprise analytics workloads.
- Built multimodal LLM pipelines with LangChain, OpenAI API, and PyTorch to support anomaly detection and RAG-based insights, improving detection accuracy by 31% and accelerating data triage across high-volume transactional and telemetry streams.
- Productionized AI microservices via FastAPI, MLflow, and Airflow, collaborating with platform, DevOps, and data engineering teams to standardize CI/CD pipelines and reduce deployment timelines by 50%.
- Implemented model optimization and monitoring using TensorRT, Prometheus, and Grafana, increasing GPU utilization by 28% while enabling real-time observability across distributed inference endpoints.
- Mentored junior engineers and partnered with data science teams to transition research models into production, cutting prototype-to-production cycles from 2 weeks to 5 days and increasing business adoption.

AI Engineer July 2024 – December 2024
VMware | Remote, USA

- Developed end-to-end AI pipelines for predictive analytics and anomaly detection using Python, PyTorch, MLflow, and Airflow, streamlining model training, validation, and deployment workflows.
- Designed scalable NLP services with Hugging Face Transformers, LangChain, FastAPI, and Docker, enabling real-time semantic search and knowledge extraction across enterprise datasets.
- Implemented model monitoring and optimization using Prometheus, Grafana, TensorRT, and Ray Serve, reducing inference latency by 27% and improving GPU utilization across distributed deployments.
- Collaborated with cloud and data engineering teams leveraging AWS (S3, EC2 GPU, Lambda), Kubernetes, and feature stores to integrate AI solutions into production pipelines while ensuring data quality, security, and reproducibility.

Data Scientist October 2020 – August 2023
Accenture | Hyderabad, India

- Led the design and deployment of enterprise AI solutions using Azure OpenAI, Hugging Face Transformers, and PyTorch Lightning, powering NLP chatbots and recommendation systems that increased user engagement by 22% and content relevance by 31%.
- Architected an end-to-end MLOps platform with Databricks, Airflow, MLflow, and Docker, automating data ingestion, training, and model promotion to reduce model delivery timelines by 45% and improve reproducibility across teams.
- Built multi-agent LLM architectures leveraging LangChain, LlamaIndex, and vector databases (Pinecone, FAISS) to enable context-aware document search and Q&A for Fortune 100 clients, cutting analyst research time by 60%.
- Collaborated with cloud and data engineering teams to modernize high-volume (5 TB/day) streaming pipelines using AWS Glue and Kinesis, integrating real-time AI services with legacy platforms and improving system throughput by 30% with zero-downtime deployments.
- Led and mentored a cross-functional team of 5 AI engineers and data scientists, establishing AI governance, observability, and coding standards that boosted team velocity by 35% and delivered 7 production-grade AI models within a year.

Core Skills

AI & Machine Learning: Python, PyTorch, TensorFlow, Scikit-learn, Hugging Face Transformers, LangChain, LlamaIndex, Open AI API, RAG, Prompt Engineering, Vector Databases (Pinecone, FAISS), XGBoost, NLP, Computer Vision, Anomaly Detection
MLOps & Infrastructure: Docker, Kubernetes, MLflow, Airflow, Databricks, Jenkins, TensorRT, Ray Serve, FastAPI, Flask, CI/C D Pipelines, Feature Store, Model Deployment, Model Monitoring, GPU Optimization
Cloud & Data Platforms: AWS (S3, EC2 GPU, Glue, Lambda, Kinesis), Azure (Data Factory, OpenAI, AKS), GCP (BigQuery, V ertex AI), Hadoop, Spark, Dask, dbt, Kafka, MongoDB, ETL Pipelines, Data Modeling
Collaboration & Leadership: Agile Methodologies, Cross-functional Teamwork, DevSecOps Collaboration, Stakeholder Engagem ent, Mentoring, Technical Documentation, AI Governance, Code Review, Knowledge Sharing

Education

MS in Data Science, Statistics, **New Jersey Institute of Technology** September 2023 – May 2025 | Newark, NJ, USA